Shubham Kumar

Noida, India
+91-8219756114 shubham.py.pyc@gmail.com Linkedin GitHub Leetcode Portfolio

Summary

GenAI Backend Engineer with 2.5 years of experience building production-grade AI systems using Python. Specialized in RAG, Graph RAG, enterprise search, and agent-based automation at scale. Experienced in designing microservices with FastAPI and gRPC, implementing vector databases and knowledge graphs, and orchestrating multi-agent workflows using LangGraph. Proven track record of delivering scalable GenAI platforms integrating LLMs, semantic search, and distributed backend systems.

EDUCATION

Chandigarh University, Gharuan, Punjab **July 2019 – Aug 2023**
Bachelor of Technology in Computer Science (CGPA: 6.97) *India*

TECHNICAL SKILLS

- **GenAI & LLM Systems:** RAG, Graph RAG, LangChain, LangGraph, LLM Orchestration, Semantic Search, Agent Systems
- **Backend & APIs:** Python, FastAPI, Flask, Django, Microservices, gRPC, REST APIs, Node.js, Express.js
- **Databases & Retrieval:** PostgreSQL, MongoDB, Neo4j, Pinecone, Qdrant, Redis
- **Cloud & DevOps:** AWS (EC2, S3), GCP, Docker, Celery, Background Workers
- **Languages:** Python, JavaScript, SQL

EXPERIENCE

Rapid Innovation **Sep 2025 – Present**
Associate Engineer – GenAI & ML Backend *Noida, India*

- Built production-grade GenAI backend systems using **RAG** and **Graph RAG** for enterprise knowledge platforms.
- Designed microservice architectures with **FastAPI** and **gRPC**, supporting scalable ingestion pipelines using **Celery** workers.
- Developed enterprise search platform integrating Gmail, GitHub, Google Drive, and Confluence using **Neo4j**, **Pinecone**, **Redis**, and **PostgreSQL**.
- Implemented multi-agent workflows using **LangGraph** and **LangSmith** with MCP-based orchestration for automation use cases.
- Optimized LLM retrieval pipelines and embedding strategies to improve response accuracy and contextual reasoning.

Avista Web Technologies **Aug 2024 – Dec 2024**
Software Engineer (Backend & AI Systems) *Noida, India*

- Designed and implemented **BPMN workflows** for Diginotary to automate digital notarization and execution tracking.
- Built and maintained **RESTful APIs** for Diginotary, DMA Guru, and HelloDR using **Node.js**, including real-time notifications with **Socket.IO**.
- Developed AI-powered backend services with **Flask** and **FastAPI**, integrating **LLM-based chatbot functionality** for healthcare use cases.
- Integrated **Stripe payments** and deployed services on **AWS (EC2, S3)** for scalable hosting and secure asset storage.

Freelance Backend Developer **Jan 2023 – July 2024**
(Python, FastAPI, Django, Flask, AWS) *Remote*

- Delivered backend systems for multiple clients using **Python frameworks** (FastAPI, Django, Flask).
- Designed REST APIs, database schemas, and scalable architectures for production applications.
- Implemented background processing, data pipelines, and cloud deployments on **AWS**.

PROJECTS

CancerGuru Platform – RAG-based Healthcare Chatbot | *FastAPI, LangChain, OpenAI* **Feb 2025 – Present**

- Built RAG-based healthcare chatbot enabling context-aware medical query responses.
- Implemented semantic search pipeline using **OpenAI embeddings** and **LangChain**.
- Designed scalable microservice deployed on **AWS**.